Requirements-Based Knowledge Discovery for Technology Management

Robert J. Watts[1], Alan L. Porter[2]
[1]U.S. Army Tank-Automotive & Armaments Command, AMSTA-TR-N, Mail Stop 272, Warren, MI 48397-5000
[2]Technology Policy & Assessment Center, School of Public Policy, and School of Industrial & Systems Engineering, Georgia Tech, Atlanta GA 30332-0205

Abstract - Exploiting information resources, particularly diverse research & development databases, presents opportunities to enhance technology management. Many such databases [e.g., *Science Citation Index, Business Index*] offer useful index fields to help you access documents. However, these indexes sometimes fail to address your requirements effectively. We present an approach that allows you to extract and categorize desired information from particular datasets that lack effective indexing for your purposes. We illustrate for a *U.S. Patents* dataset on lightweight automotive materials.

INTRODUCTION

Managing technology presents intriguing challenges and opportunities in an era marked by accelerating rates of technical change, increasing cross-organizational competition and collaboration, and expanding information resources [8] [13]. Anticipating the directions and implications of technological change presents corresponding challenges and opportunities "in press"[2]. Electronic abstract databases on research and development (R&D) offer a critical set of information resources for those needing technological intelligence. But, how can one effectively exploit such resources?

"Text mining" describes one set of tools to help analysts extract valuable intelligence from massive volumes of R&D descriptions – e.g., thousands of abstracts on some topic of interest. Convergent approaches include:

- "KDD" – Knowledge Discovery in Databases [21][22]
- "scientometrics" [3][11][16][17]
- "co-citation analysis" [5]
- "text data mining" [9]
- "technology opportunities analysis" [12][15]

These approaches can help technology managers answer varied questions about an emerging technology, for instance:

- Our researchers are intrigued with a particular technological development pathway – what are our competitors doing in this area?
- Should we pursue this technology, or is it so well protected by others' intellectual property rights that this would be foolhardy?
- Where might we apply such technologies – what industrial sectors evidence interest?
- Are there more promising candidate technologies to fulfill the target needs?

Answering such questions poses challenges. Some databases are well indexed (e.g., *MEDLINE*, covering the world's medical R&D quite comprehensively), but those well-structured indexes may be slow to capture rapidly emerging science and technology. Other databases are notoriously poorly indexed (e.g., *U.S. Patents*, lacking even suitable "keywords" to help locate relevant patents amidst abstracts often written to obfuscate intended applications).

Once one has separated out the "signal" from the "noise," one faces challenges in extracting answers effectively. Different analyses answer different questions. Different representations convey different information. Busy technology managers often challenge the credibility of text mining to extract meaningful results from large databases of information. Managers are likely to be naïve about the

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
29 JUL 2001	**N/A**	

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Requirements-Based Knowledge Discovery for Technology Management	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
; Watts /Robert,JPorter /Alan,L	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
US Army RDECOM-TARDEC 6501 E 11 Mile Rd Warren, MI 48397-5000	**17073**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
TACOM TARDEC	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release, distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**SAR**	**13**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

information resources (databases), ill at ease with statistical techniques, and hard-pressed to relate these new forms of knowledge to their pressing issues. From the analyst's perspective, complementary challenges loom – it is hard to communicate complex relationships quickly.

This paper tackles these information-mining issues. It focuses on a realistic technology management domain – advanced materials for use in automotive applications (of concern to the Army to develop future military vehicles). We perform searches on the same topic in two R&D publication abstract databases -- *EI Compendex* (*Engineering Index*) and *INSPEC*. These databases provide useful descriptors to help categorize issues concerning lightweight materials. We use *Tech OASIS/VantagePoint* text mining software's list comparison, natural language processing, and principal components decomposition features to apply these descriptors to and categorize the patents. This provides a much more informative profile of this competitive technology. We discuss how various descriptor sets, such as internal requirements lists, can be used to generate managerial value from extensive information resources. Within this domain, we focus down on particular lightweight materials to illustrate how certain text mining methods can help extract valuable information from almost impenetrable resources (huge patent search sets). It does so from the point of view of highly knowledgeable professionals and managers exploring the potential of a relatively uncommon automotive material – magnesium. (Technology management questions and appropriate ways to answer them would differ considerably for newcomers exploring such a topic, or again for patent attorneys assessing intellectual property prospects – techniques must be adapted to provide the needed information.)

We abstract from this broad technology intelligence problem to illustrative particulars. Our intents are to introduce the readers to

- The general notion of treating large bibliographic datasets
- Using multiple datasets to understand a topic better
- A novel way to inductively "index" an unindexed search set
- "Text mining" technological literature to help manage technology.

We must warn the reader that some elements of this discussion will likely feel rather technically dense. If that happens, we ask you to skim forward to get the gist of the approach and imagine how such techniques might be adapted to your own information needs concerning trends and relationships among emerging technologies.

METHODS

Some R&D literature abstract databases, such as *INSPEC* or *EI Compendex*,[1] contain fields that reflect the contents of the associated documents – we call these "keywords." These may be tightly controlled subject index terms and/or author-generated descriptors. Other databases, such as *U.S. Patents*, lack such helpful descriptors.

In seeking to profile R&D in a target domain, one might search in such databases. The search results can be retrieved electronically quite easily these days. Arrangements can be made with database suppliers for licenses to download unlimited numbers of records over a chosen time period (typically a year). In our example analysis, we retrieved 2185 patent abstracts relating to "lightweight automotive materials." One issue of interest to Army technology managers is to find out how magnesium might be applied in future automotive applications. Very few of the patents explicitly mention magnesium. Hence, we wanted to explore how other lightweight technologies might relate to particular automotive functions from which we could extrapolate potential magnesium applications.

Analytical tools are being developed to help extract valuable information from such text resources. These entail counting – known as "bibliometrics" (e.g., which organizations are securing the most lightweight materials patents? what materials are most frequently mentioned in conjunction with

[1] *INSPEC* is produced by IEE and *EI Compendex* by Engineering Information (Elsevier). They each cover substantial segments of engineering R&D journal articles and conference papers.

automotive applications?). Furthermore, software can help probe texts more deeply. We use a program known commercially as *VantagePoint* and as *Tech OASIS* in the Department of Defense.[2] This brings to bear statistical analyses enhanced by natural language processing (NLP), computational linguistics, and fuzzy matching. It is especially suited to analyze literature abstracts (i.e., field-delineated records). *Tech OASIS* provides a capability to tally counts (lists) and to explore linkages based on **co-occurrences of terms** across the abstract records (matrices). Analyses can be based on a complete search set of abstracts or on subsets of interest.

Tech OASIS provides numerous capabilities that can be brought to bear to answer particular questions (c.f., http://theVantagePoint.com). Under the rubric, "Technology Opportunities Analysis," our research group has explored how these can help resolve various technology management information needs (c.f., http://tpac.gatech.edu). The authors have used *Tech OASIS* in developing "innovation forecasting" [20]. Innovation forecasting applies conceptual models related to the processes of technology substitution, diffusion, and transfer to measure factors (i.e., technological, industry infrastructure, economic, socio-political, and institutional) that should be weighed when forecasting technological innovation.

One subset of text mining techniques concentrates on helping users sift through huge sets of abstract records to find subsets that concentrate on particular issues of special interest. Techniques exist to logically group individual documents based on high co-occurrence patterns of each record's "keywords" (or title phrases, etc.) with the other analyzed records' keywords. *Tech OASIS* groups such related records based on principal components analysis (PCA – a form of factor analysis – [4][6]). We have developed further proprietary analysis techniques, based on PCA, to inductively form related document groupings. Our "broad base principal components decomposition" ("BB-PCD") uses an iterative optimization process to create record groups with maximally shared interests. It strives to include as many of the records as possible in an optimal number of groups (factors), while minimizing duplication among groups [19].

Fig. 1 provides a simple example of a partial hierarchy diagram for 82 R&D publication abstracts from the automotive lightweight materials search. These 82 abstracts were all published since 1994 and contain the term "magnesium." Normally, principal components analysis, as well as BB-PCD, would be applied to larger abstract sets. Our specific use of these abstracts will be elaborated later.

The hierarchy of Fig. 1 reflects an initial decomposition into four groups – shown as boxes on the top row. The terms shown in each box are the most related keywords that tend to occur together, forming that group of records. The key advantage of this approach is that it avoids pre-established index structures, instead reflecting the actual relationships inherent in these abstract records. Note that the third box contains the greatest number of the 82 records; it appears to relate to both automotive applications and metals. Underneath this box are shown a second-tier breakout of groupings within this set of 54 records (we could show similar breakouts for the other boxes). For instance, the third group down seems to relate to "industrial machining of magnesium castings." Further examination of this focused subset may prove valuable in discovering new topical relationships. Linkage to the individual records is maintained throughout the *Tech OASIS* analysis processes so one can "dig down" to explore groups of interest. To reiterate, this hierarchical decomposition shows groups of abstract records that use particular terms together more frequently than one would expect based on raw term frequencies. Some such groupings prove to be noise. But others elucidate non-obvious interrelations that can be amazingly informative in understanding topical relationships. One can also identify which research organizations are engaging particular related topics -- perhaps to identify potential collaborators from seemingly unrelated disciplines [20].

---

[2] The U.S. Army Tank-automotive and Armament Command's National Automotive Center's (NAC) mission includes promoting and establishing collaborative programs between academia, industry and government agencies. Like purely industrial partnering, these collaborations strive to address common needs and intend to promote efficient R&D resource utilization. More importantly, the government collaboration process strives to strengthen technological and industrial infrastructures, making them better prepared to meet U.S. military material requirements in a timely and cost efficient manner. Under this mission, the NAC has collaborated with the Defense Advanced Research Projects Agency (DARPA) in the development of the Technology Opportunities Analysis of Scientific Information System (*Tech OASIS*), a user-friendly software system that aids in literature research. The *Tech OASIS* development effort represents a collaborative partnership itself, which includes Search Technology, Inc. as the prime contractor and sub-contractors, the Georgia Tech Technology Policy and Assessment Center (TPAC) and Intelligent Information Services Corporation (IISC).

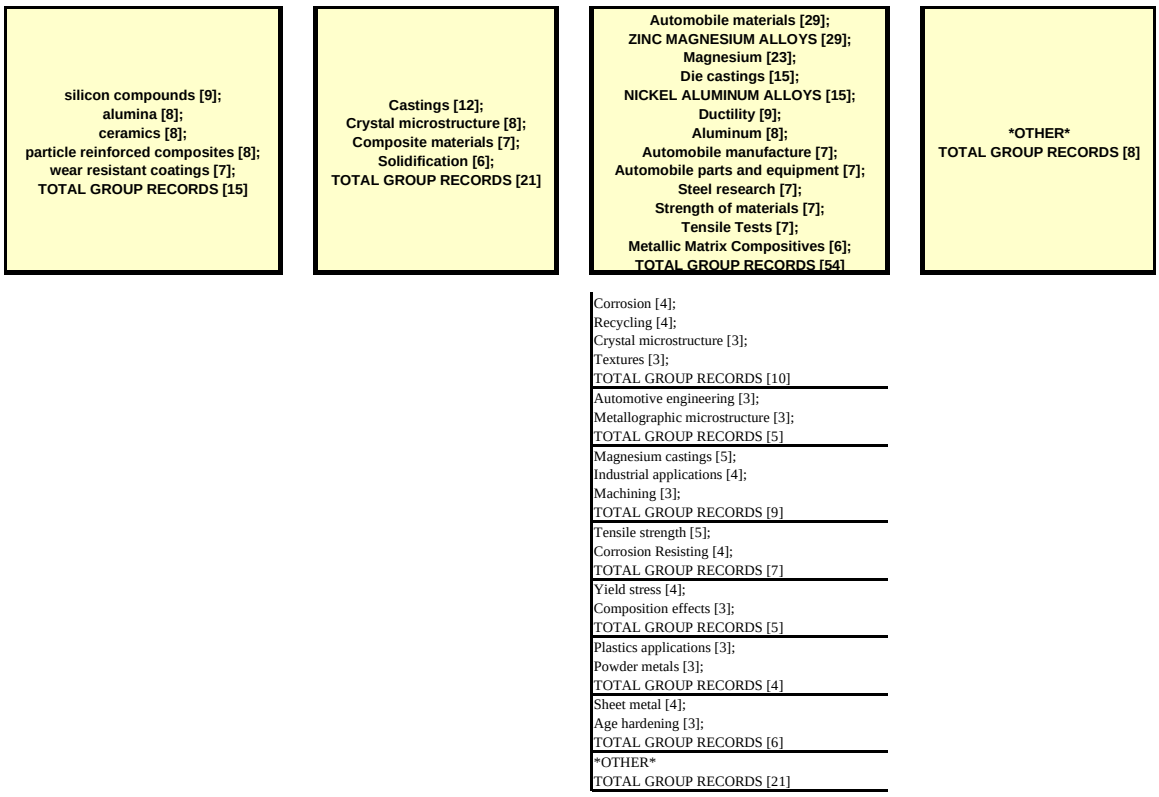

Fig. 1 Hierarchy for 82 R&D Automotive Lightweight Materials Abstracts Containing "Magnesium"

USING ONE DATABASE TO SHED LIGHT ON ANOTHER

Difficulties often arise when retrieving information from databases comprised of document abstracts [10]. Databases such as *U.S. Patents* lack a field that reflects each document's contents (e.g., keywords).[3] In other cases, keyword analyses may not generate groupings that embody the focus area of interest. Extracting a suitable set of keywords from one dataset (e.g., R&D publications) can serve to elucidate the structure of another dataset (e.g., patents).

Table 1 provides the "Descriptors" (i.e., keywords) occurring 3 or more times in the 82 records of Fig. 1. General, non-descript, terms, such as those highlighted in Table 1, can muddy analyses. Beginning with a set of more sharply targeted terms can enhance the resultant groupings. One way to improve lists is to look across keywords from multiple searches, drawn on the same topic, but from different databases. Different databases emphasize different aspects. For instance *Business Index* spotlights business aspects, whereas *EI Compendex* focuses on engineering issues, as does *INSPEC*, but with more emphasis on research and certain topics (e.g., electronics, computing).

The described application draws on patent abstracts. *U.S Patents* abstracts do not have a suitable keywords field. Using the *Tech OASIS* natural language processing (NLP) capability, noun phrases can be extracted from the abstract (or title) fields. The NLP-extracted nouns/noun phrases, however, not only contain the general, non-descript type of term/phrases discussed in the last section, but also contain legalistic terms/phrases. Table 2 provides the 75 most frequently occurring phrases from the 2185 "automotive lightweight materials" patents. Illustrative unhelpful noun phrases have been highlighted.

[3] For these statistical analyses it is also essential that such fields include multiple values to characterize each document. *Tech OASIS* can parse the abstracts themselves into noun phrases, but often these are too low frequency to yield effective BB-PCD groupings.

TABLE 1
DESCRIPTORS-ABSTRACTS ON AUTOMOTIVE LIGHTWEIGHT MATERIALS & MAGNESIUM

#	Descriptors (Cleaned)	#	Descriptors (Cleaned)	#	Descriptors (Cleaned)
31	magnesium alloys	7	wear resistant coatings	4	Recycling
29	Automobile materials	6	copper alloys	4	reviews
29	ZINC MAGNESIUM ALLOYS	6	Corrosion	4	Sheet metal
23	Magnesium	6	Metallic Matrix Compositives	4	Yield stress
21	aluminium alloys	6	Solidification	3	Age hardening
15	Die castings	6	surface topography measurement	3	Automotive engineering
15	NICKEL ALUMINUM ALLOYS	6	yield strength	3	chromium alloys
11	silicon alloys	5	elongation	3	Composition
10	casting	5	Extrusion	3	Composition effects
9	Ductility	5	Hardness	3	cracks
9	silicon compounds	5	Magnesium castings	3	Creep
8	alumina	5	precipitation hardening	3	Elastic moduli
8	Aluminum	5	Tensile Tests	3	Fatigue testing
8	ceramics	5	zinc alloys	3	forming processes
8	Crystal microstructure	4	Adhesion	3	fracture toughness
8	particle reinforced composites	4	ageing	3	laser beams
8	scanning-transmission electron microscopy	4	Corrosion Resisting	3	Metallographic microstructure
8	Tensile strength	4	fibre reinforced composites	3	Microhardness
7	automobile industry	4	Forgings	3	Mixing
7	Automobile manufacture	4	Industrial applications	3	Plastics applications
7	Automobile parts and equipment	4	Machining	3	Powder metals
7	Composite materials	4	Magnesium compounds	3	Protective coatings
7	Steel research	4	nickel alloys	3	Scanning electron microscopy
7	Strength of materials	4	Porosity	3	transmission electron microscopy

TABLE 2
ABSTRACT NOUN PHRASES FROM "AUTOMOTIVE LIGHTWEIGHT MATERIALS" PATENTS

#	Descriptors (Cleaned)	#	Descriptors (Cleaned)	#	Descriptors (Cleaned)
376	materials	112	least one pocket	72	members
341	automobiles	112	sides	72	panels
334	uses	107	plastics	72	temperatures
317	methods	105	substrates	71	bases
252	inventive	103	heating	71	products
230	one	101	components	69	manufacturing
225	vehicles	100	molds	69	polymers
223	plurality	100	weight	68	shaping
210	surfaces	95	housings	65	elements
209	restrained automotive vehicle	91	coats	65	windows
160	processing	89	applicator	60	lights
153	structures	86	body	60	pressurized
152	present invention lies	83	mixtures	59	opposed inner surface
150	portions	81	adhesives	58	metals
145	compositions	81	systems	57	frames
143	least	79	covers	57	front
137	first	79	transparent plastic material	57	water
134	opens	78	B	56	edging
132	devices	77	forms	55	amounts
127	pairs	77	smooth outer surface	55	one side defining
125	parts	75	cS	55	Zn sub 2
121	disclosed	74	positions	54	1
116	assembly	74	sheets	54	air
114	apparatus	73	glasses	53	one tapered end
113	ends	72	contacts	53	seals

One approach to sift through keywords is to use thesauri that tag sets of terms of special interest (e.g., "materials," "engine applications") [1]. If one had such a thesaurus for, say, plastics, it could sort through the terms in Table 2 to tag related terms: plastics, transparent plastic material, polymers, etc. We lacked thesauri containing "automotive lightweight material" terms and phrases in doing the current analyses. To simulate such a thesaurus, we draw on keywords from publication database searches to help extract groups of patent records (the idea of drawing on an alternative resource to generate a thesaurus is not new – [7]).

Literature searches on "lightweight automotive materials" were conducted in the *INSPEC, EI Compendex* and *U.S. Patents* databases (Table 3). Line "S15" shows the final search string used. Resulting R&D abstracts from the *INSPEC* (619 records) and *EI Compendex* (3458 records) search sets were combined. Then, the abstracts that were published since 1994 and that contained the term "magnesium" were separated into a new dataset, which contained 82 R&D literature abstracts (recall Table 1). Keywords from these 82 publication abstracts serve as our thesaurus list. Note that the terms in Table 1 are technically richer than those in Table 2.

TABLE 3
AUTOMOTIVE LIGHTWEIGHT MATERIALS LITERATURE SEARCH STRATEGY

		INSPEC	EI Compendex	U.S. Patents
Set	Description	Items	Items	Items
S1	(AUTOMOTIVE OR AUTOMOBILE) AND MATERIAL?	1140	12,027	3441
S2	S1 AND PY>1989	1100	5006	File 654
S3	RD S2 (unique items)	1076	4827	NA
S4	S3 AND (PLASTIC? OR POLYMER?)	190	1307	1130
S5	S3 AND COMPOSITE?	133	725	230
S6	S3 AND ALUMINUM	59	665	272
S7	S3 AND TITANIUM	46	113	89
S8	S3 AND ALLOY?	186	822	188
S9	S3 AND CERAMIC?	101	380	148
S10	S3 AND STEEL	90	990	222
S11	S3 AND GLASS	45	294	400
S12	S3 AND (STRUCTURE? OR STRUCTURAL)	166	941	794
S13	S3 AND ((FINITE()ELEMENT()ANALYSIS) OR FEA)	71	94	0
S14	S3 AND METALLIC	34	203	128
S15	S4 OR S5 OR S6 OR S7 OR S8 OR S9 OR S10 OR S11 OR S12 OR S13 OR S14	619	3458	2185

Patent Search Limited to the Fields:Title,Abstract & Claims

The resulting keyword set emphasizes research terms. Depending on the questions one seeks to answer, other keyword sets might be constructed to get at different product value chain functions (e.g., to seek complimentary technologies, processing materials, manufacturing and machining approaches, product functions).

The complete list of keywords (not shown in Table 1) from the combined *INSPEC* & *EI Compendex* file (i.e., our 82 R&D record-based set) was compared to noun phrases from the 2185 "automotive lightweight materials" patent abstracts (Table 3). The common noun phrases were then used as inputs to the BB-PCD analysis process. The resulting groupings of patent abstracts prove of interest.

The titles for the derived record groupings are shown as the column headers in the third row of Table 4. Above the group names, in the second row of Table 4, are the number of records grouped in each category. Shown in the second column of Table 4 are the abstract noun phrases that were common to our surrogate thesaurus term list and determined to be group-defining terms by the BB-PCD analysis process. The X's in each column designate the group-defining noun phrases for the respective group. For instance,

the abstract noun phrases “aluminum,” “copper,” and “magnesium” define the “aluminum” group of 50 patent abstracts.

TABLE 4
AUTOMOTIVE LIGHTWEIGHT MATERIALS PATENTS BB-PCD GROUPS VS. GROUP DEFINING NOUN PHRASES

		List Compare BBPCD Derived Record Categories																	
	# Records	958	50	21	58	33	317	220	25	44	15	17	39	21	33	341	159	225	107
# Records	Abstract Noun Phrases	*OTHER*	aluminum	elasticity	thermoplastics	castings	processing	surfaces	thermal expansion	cutting	friction	stresses	mechanical properties	tensile strength	steel	automobiles	assembly	vehicles	plastics
341	automobiles	Not														X			
225	vehicles	Not																X	
210	surfaces	Not						X											
160	processing	Not					X												
153	structures	Not					X												
116	assembly	Not															X		
107	plastics	Not																	X
48	thermoplastics	Not			X														
40	aluminum	Not	X																
34	alloys	Not															X		
33	catalysts	Not					X												
30	cutting	Not								X									
22	steel	Not													X				
19	castings	Not				X													
18	mechanical properties	Not											X						
17	adhesion	Not						X											
17	ceramics	Not				X													
17	stresses	Not										X							
15	automotive industry	Not								X									
15	friction	Not									X								
15	thermal expansion	Not							X										
14	durability	Not															X		
14	room temperature	Not											X						
14	strength	Not													X				
13	mixing	Not			X														
13	tensile strength	Not												X					
12	elasticity	Not		X															
11	automobile industry	Not		X															
11	copper	Not	X																
11	corrosion	Not											X						
11	joining	Not							X										
11	magnesium	Not	X																
11	solution treatment	Not												X					

Overall, this list comparison process grouped 1227 of the 2185 patents. Analysts and end-users can examine these groups to better understand lightweight automotive materials patenting. The resulting groups reflect empirical relationships, not preordained ideas of what goes with what.

However, human review and focused value assessment of 1227 patent abstracts would be both time-consuming and difficult. This is particularly so given our focus – what roles might magnesium play in automotive design? Hence we use an iterative analysis process to winnow down the 1227 abstracts to more sharply defined groups.

In this example, we have defined our focus area of interest as "magnesium" within the broader area of R&D on "automotive lightweight materials." To create and designate the focus area, we manually create a group containing any of four noun phrases (magnesium, magnesium alloy, least magnesium, and fibre-reinforced magnesium alloy). These four magnesium terms are contained in 15 "on-target" patents. The group of magnesium noun phrases serves as the technology focus area.

We based the initial decomposition of the 2185 patents on the noun phrases in common with the keywords from the 82 publication abstracts. We now generate a co-occurrence matrix between those derived BB-PCD groups and the focus group terms (the four magnesium phrases from the selected 15 patents). Co-occurrence counts determine the groups for file recombination.

In the first iteration portion of Table 5, the records from groups that have a negligible count of the focus group's terms are eliminated from the file. Thus, the records from the following groups are eliminated: OTHER, automobiles, vehicles, surfaces, thermoplastics, cutting, steel, thermal expansion, elasticity, stresses, and friction. The remaining 382 records are analyzed during the second iteration, after which the records from groups: structures, solution treatment, and OTHER, are eliminated from the analysis. The remaining 268 records are categorized into seven BB-PCD groups, all of which include focus group terms.

The 2185 patents have thus been filtered down to 268 abstracts, and those have been grouped into seven categories. The aluminum group has the greatest density of the focus group terms -- 10 of 27 or 37% of these patents contain the term magnesium. New knowledge may be contained in the other 17 patent abstracts. These are our "plums" – patents that don't explicitly mention magnesium, but have much in common with patent abstracts that do address magnesium. Expert review is required to make this determination.

Similarly, the technology focus group record densities of the remaining six groups (Third Iteration, Table 5) can guide detailed expert review. In descending order, we might explore alloys (12%), mechanical properties (5%), castings (4%), processing (3.3%), assembly (3.2%) and plastics (3.1%). The technical expert might also use these technology group-defining terms to further guide detailed patent review.

We should note that the patent exploration process is not "done." The analyses have focused energies on particular prime groups of the 2185 patent abstracts. When review of those abstracts identifies particularly interesting patents, the analyst and end-user may well decide to retrieve the full patent records (not available in this *U.S. Patents* database search set). But instead of paying for and being buried under 2185 full patents, we concentrate our energies on a precious few.

Tech OASIS also provides visualization/mapping capabilities. Fig. 2 is a cross-correlation map of the third iteration's seven derived BB-PCD groups. Positioning (using Multi-Dimensional Scaling) and linkages (using a special Path-Erasing Algorithm) are based on co-occurrence of the focal keywords. Solid, dashed, and no connecting lines indicate the three levels of relatedness among the nodes. Larger nodes represent larger groups (more records).

Five of the groups in Fig. 2 -- aluminum, mechanical properties, castings, alloys and processing -- are highly related to one another. The remaining two groups -- assembly and plastics -- are moderately related to each other and low in relation to the other five groups. Assuming that the expert interests lie in the central, five-group, body of knowledge, the patent review could then be focused on 158 patents contained therein. Thirteen of these specifically use the magnesium focus terms/phrases. The remaining 145 patents were extracted and categorized based upon the abstracts' term content common to the technology focus terms (the four magnesium-related phrases). Theoretically, these patents should embody related product value chain technologies/information for magnesium automotive lightweight materials. Their significance must be judged by expert review.

TABLE 5
THREE ITERATIONS; BB-PCD GROUPS VS. FOCUS GROUP TERMS CO-OCCURRENCES

	# Records	11	3	1	1
# Records	**First Iteration**	magnesium	magnesium alloy	least magnesium	fiber reinforced magnesium alloy
2185	Total Records	11	3	1	1
958	*OTHER*				
341	automobiles	1			
317	processing	4			
225	vehicles				
220	surfaces				
159	assembly	3		1	
107	plastics	2			
58	thermoplastics				
50	aluminum	11			
44	cutting				
39	mechanical properties		1		1
33	castings		1		
33	steel				
25	thermal expansion				
21	elasticity				
21	tensile strength	1	2		
17	stresses				
15	friction				

	# Records	10	3	1	1
# Records	**Second Iteration**	magnesium	magnesium alloy	least magnesium	fiber reinforced magnesium alloy
382	Total Records	10	3	1	1
112	PCD: structures				
104	PCD: processing	3			
68	PCD: assembly	1		1	
66	PCD: plastics	2			
29	PCD: alloys	2			
27	PCD: aluminum	10			
23	PCD: castings		1		
20	PCD: mechanical properties		1		1
11	PCD: solution treatment	1	2		
2	PCD: *OTHER*				

	# Records	10	3	1	1
# Records	**Third Iteration**	magnesium	magnesium alloy	least magnesium	fiber reinforced magnesium alloy
268	Total Records	10	3	1	1
90	PCD: processing	3			
65	PCD: plastics	2			
62	PCD: assembly	1		1	
34	PCD: alloys	2	2		
27	PCD: aluminum	10			
24	PCD: castings		1		
19	PCD: mechanical properties		1		1

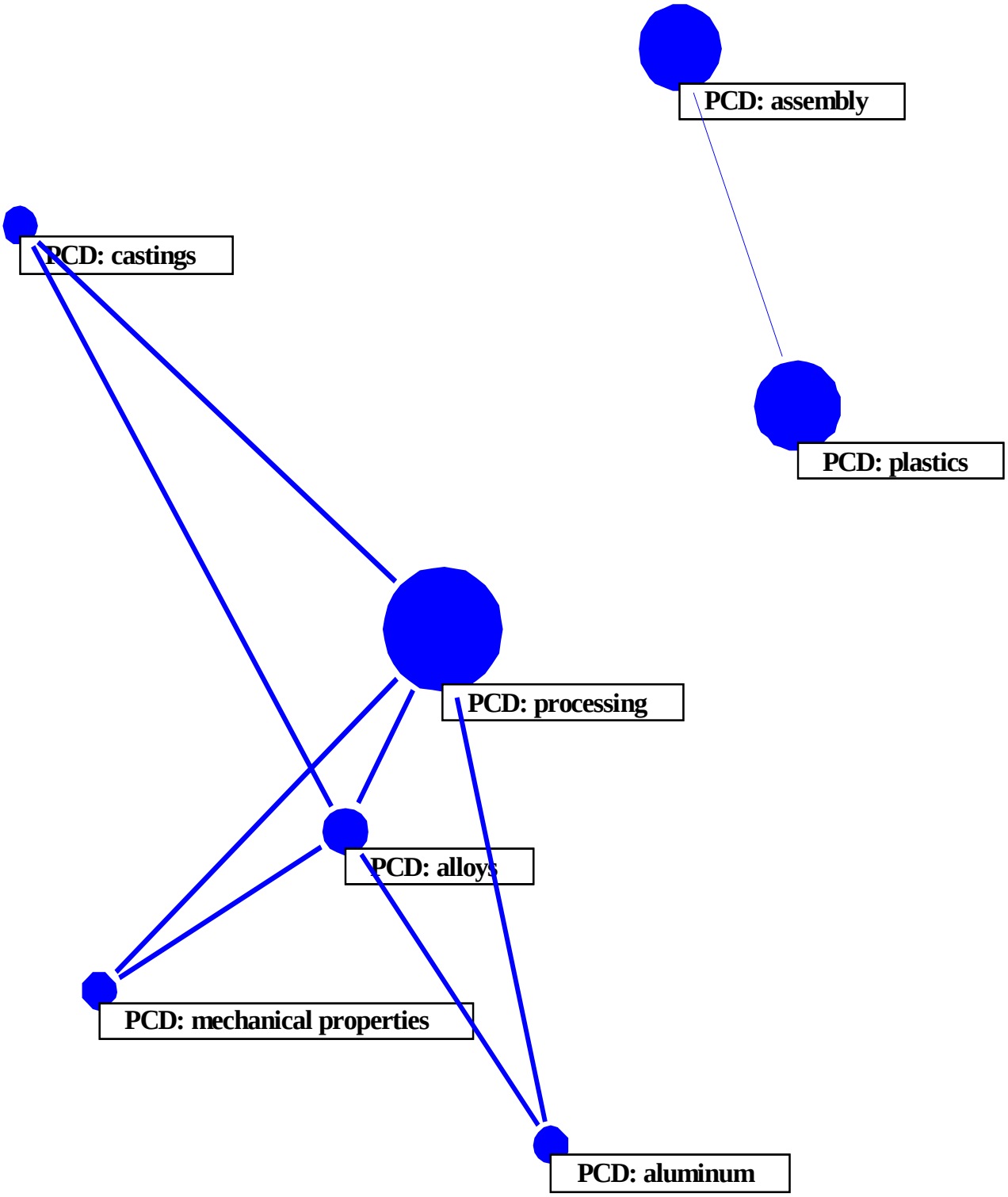

Fig. 2. Magnesium PCD Focus Groups in Automotive Lightweight Materials

We note that interests sometimes lie not in the central topical area but along the "fringes." We have devised another analytical process called "discrete segment principal components decomposition" to help pull out the marginally related records [19]. In Fig. 2 one could imagine pursuing, say, "plastics" for creative ideas on how to apply magnesium. The expectation would be that most technologies, processes, and applications suggested in such exploratory investigation would not be meaningful. However, the aim is "discovery" of remote links and possibilities. Interesting cross-disciplinary links have been uncovered through such text mining [14][9]. In essence, the logic is that A (e.g., magnesium) has some commonalties with B (e.g., plastics). B has some commonalties with C (e.g., automotive applications). Consider the possibility of A-C association.

DISCUSSION

We have illustrated a text decomposition process that applies two lists of terms/phrases related to the analyzed body of information. First, a set of domain specific terms (a thesaurus) identifies the portion of information (i.e., abstracts or records) to be logically categorized. In this case, such terms were derived from a composite of two R&D publication abstract sets pertaining to magnesium, then applied to a patent abstracts set. Using *Tech OASIS*, the resulting groups of patents provided some structure to a large set of abstracts relating to lightweight automotive materials.

However, we wanted to further filter the resulting 1227 patents. We devised a small set of magnesium-related patent abstracts to provide technological focus terms. Then we used those terms in an iterative process to form more focused patent groups relating to magnesium (but not restricted to explicit use of "magnesium" per se). At the time of writing of this paper, subject matter expert review was being sought to assess the relevance of the derived groupings and the associated embodied information. (This should be complete by the time of the conference.)

Several potential knowledge discovery applications exist for the subject-focused decomposition process presented. As demonstrated, the algorithm permits analysis of databases that lack keywords that reflect documents' contents (e.g., for patent analyses).

Another application could entail searching on affiliation, followed by analyses on one or more technological dimensions. For instance, one might retrieve all the patents of a given company. Then, one could group those patents according to a particular special set of terms (thesaurus) relating to a technology domain.

A third tack would associate distinct text sources. For instance, the U.S. Navy annually identifies technological requirements for some 100 critical technologies. These are presented as text descriptions reflecting a composite of capabilities sought. Natural language processing of these short texts can provide a well-targeted set of terms (thesaurus). That could then be applied to other text resources to identify potential links. Those other text resources could be external technology R&D databases such as used in this paper, or trip reports, internal R&D project descriptions, or Internet site contents.

As a final illustration, the process could aid large organizations' program management oversight. Organizations performing a significant amount of dispersed research and development face a challenge in cross-fertilizing related projects. For this application, the records' file analyzed would be of individual program descriptions and progress reports. The thesauri might contain the desired technology thrusts and focus areas for the corporation or agency.

Our future efforts will concentrate on developing approaches for creating, obtaining, and improving the thesauri (term or phrase lists) used in the subject-focused decomposition process. Also, we must strive to create a thesaurus of generic/non-descript terms and phrases, which could be used to exclude such terms/phrases from the analysis. Focused applications must be performed. We have observed that subject matter experts can quickly observe and/or identify a group of terms or phrases that define a focused area of R&D. Therefore, subject matter expert involvement must be obtained for more thorough algorithm development and applications.

In addition we mention the ongoing development of *Tech OASIS/VantagePoint*. The aim is to increase the power of such text mining tools and, more importantly, their usability. Toward this end, we note efforts to:

- provide a "smart front-end" (to enable access to multiple information resources without having to learn every protocol and to facilitate rapid search refinement)
- develop scripts (VisualBasic macro's) that automate the generation of effective information products (e.g., when an end-user identifies an effective set of charts on one technology analysis to enable just-in-time replication on other technologies)
- enhance visualization of trends and relationships
- determine what various technology managers find most effective – there are major obstacles to utilization of information mining tools – and hasten their application to derive useful technology intelligence.

Exploiting external information resources offers a critical edge to technology managers. Combined with expert opinion, analyses of publication, patent, citation, and project databases can provide essential competitive technological intelligence. Increasingly powerful software tools to help extract vital insights. Our forecast: technology managers who fail to utilize such tools will be supplanted by those who do by 2010. Why? Better-informed decisions about technology priorities and investments will win out over intuitive ones.

REFERENCES

[1] Chen, H., Lynch, K.J., Koushik, B., and Ng, T.D., "Generating, Integrating, and Activating Thesauri for Concept-based Document Retrieval," *IEEE Expert*, p. 25-34, 1993 (April).

[2] Coates, V.T., Farooque, M., Klavans, R., Lapid, K., Linstone, H.A., Pistorius, C., and Porter, A.L., "The Future of Technological Forecasting," *Technological Forecasting and Social Change*, to appear.

[3] *Current Science* (Vol. 79, No. 5, 10 Sep., 2000), special section on *Scientometrics*: http://ces.iisc.ernet.in/Current_Science/

[4] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, D., (1990), "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science* 41(6), p. 391-407, 1990.

[5] Franklin, J.J. and Johnson, R., "Co-citation Bibliometric Modeling as a Tool for S&T Policy and R&D Management: Issues, Applications, and Developments," in *Handbook of Quantitative Studies of Science and Technology*, A.F. J. van Raan, ed., Elsevier: Amsterdam, 1989.

[6] Harman, H. H., *Modern Factor Analysis*, University of Chicago Press: Chicago, 1967.

[7] Hwang, D., and Nagao, M., "Construction of a thesaurus for Korean from a thesaurus for Japanese," *Transactions of the Information Processing Society of Japan* 35(2), p.210-21, 1994.

[8] Kash, D.E., and Rycroft, R.W., "Patterns of Innovating Complex Technologies: A Framework for Adaptive Network Strategies," *Research Policy* 29, p. 819-831, 2000.

[9] Kostoff, R.N., and Geisler, E. "Strategic Management and Implementation of Textual Data Mining in Government Organizations," *Technology Analysis & Strategic Management* 11(4), p. 493-525, 1999. See also Kostoff websites: http://www.dtic.mil/dtic/kostoff/index.html; and http://www.sciquest.com/cgi-bin/ncommerce3/ExecMacro/sci_ethics.d2w/

[10] Mauldin, M.E., *Conceptual Information Retrieval: A Case Study in Adaptive Partial Parsing*, Kluwer, Boston, 1991.

[11] Narin, F., Loivastro, D., and Stevens, K.A., "Bibliometrics -- Theory, Practice and Problems," *Evaluation Research* 18(1), 65-76, 1994.

[12] Porter, A.L. and Detampel, M.J., "Technology Opportunities Analysis," *Technological Forecasting & Social Change* 49, p. 237-255, 1995

[13] Rycroft, R.W., and Kash, D.E.: *The Complexity Challenge: Technological Innovation for the 21st Century*. Pinter, London, UK, 1999.

[14] Swanson, D.R., and Smalheiser, N.R., "An Interactive System for Finding Complementary Literatures: A Stimulus to Scientific Discovery," *Artificial Intelligence* 91(2), p. 183-, 1997.

[15] Technology Policy and Assessment Center, Georgia Tech: see "Technology Opportunities Analysis," and "HotTech" on http://tpac.gatech.edu

[16] Van Raan, A.F. J., ed., *Handbook of Quantitative Studies of Science and Technology*, Elsevier: Amsterdam, 1989.

[17] van Raan, A. (and colleagues at the University of Leiden) – bibliometric mapping: http://sahara.fsw.leidenuniv.nl/cwts/

[18] *VantagePoint* is described at: http://theVantagePoint.com/

[19] Watts, R.J., Porter, A.L., and Courseault, C., "Functional Analysis: Deriving Systems Knowledge form Bibliographic Information Resources, *Information, Knowledge, Systems Management* 1(1), p. 45-61, 1999.

[20] Watts, R.J., and Porter, A.L., "Innovation Forecasting," *Technological Forecasting and Social Change* 56, p. 25-47, 1997.

[21] Watts, R.J., Porter, A.L., Cunningham, S. & Zhu, D., "TOAS Intelligence Mining; Analysis of Natural Language Processing and Computational Linguistics," in *Principles of Data Mining and Knowledge Discovery*, J. Komorowski and J. Zytkow (eds.), (First European Symposium -- *PKDD'97, Trondheim, Norway)*, p. 323-335: Springer, 1997.

[22] Zhu, D., Porter, A.L., Cunningham, S.W., Carlisle, J., and Nayak, A., "A Process for Mining Science & Technology Documents Databases, illustrated for the case of "Knowledge Discovery and Data Mining," *Ciencia da Informacao* 28(1), p. 1-8, 1999.